Discrete Random Structures in Bayesian Nonparametrics

II: Models beyond the Dirichlet process

Igor Prünster

Bocconi University

Outline

Prediction beyond the DP & Gibbs-type framework

Problem: How does one obtain novel families of predictive distributions i.e. beyond the DP & the Gibbs-type framework?

Recall that within the exchangeability framework one has:

- Singular (or one–step) prediction

$$\underbrace{\mathbb{P}[X_{n+1} \in A \,|X^{(n)}]}_{\text{predicitve distribution} \in \mathscr{P}} = \int_{\mathscr{P}} P(A) \underbrace{Q(\mathrm{d}P|X^{(n)})}_{\text{posterior distribution}}$$

- m–step prediction

$$\mathbb{P}[X_{n+1} \in A_1, \ldots, X_{n+m} \in A_m|X^{(n)}] = \int_{\mathscr{P}} \prod_{i=1}^{m} P(A_i)\, Q(\mathrm{d}P|X^{(n)}).$$

Solution: There are two possible strategies to obtain novel families of predictive distributions

1. Direct approach: specify a system of conditional distributions in a way that the exchangeability assumption is preserved.
2. Indirect approach: use de Finetti's representation theorem in an instrumental way.

Direct approach to prediction

Goal: define a collection $(p_n)_{n\geq 1}$ of transition kernels on $\mathbb{X} \times \mathbb{X}^n$

$$p_n(A; \boldsymbol{x}^{(n)}) = \mathbb{P}[X_{n+1} \in A \,|\, X_1 = x_1, \ldots, X_n = x_n]$$

in such a way that the sequence $(X_n)_{n\geq 1}$ is exchangeable.

Answer: Characterization in Fortini, Ladelli & Regazzini (2000)
A sequence of kernel transitions $(p_n)_{n\geq 1}$ identifies the law of an exchangeable sequence $(X_n)_{n\geq 1}$ if and only if

(i) *For any permutation π of $\{1, \ldots, n\}$*

$$\mathbb{P}[X_{n+1} \in A \,|\, X_1 = x_1, \ldots, X_n = x_n] = \mathbb{P}[X_{n+1} \in A \,|\, X_{\pi(1)} = x_{\pi(1)}, \ldots, X_{\pi(n)} = x_{\pi(n)}]$$

(ii) *Two step ahead predictive invariance*

$$\mathbb{P}[X_{n+1} \in A,\, X_{n+2} \in B \,|\, X_1, \ldots, X_n] = \mathbb{P}[X_{n+1} \in B,\, X_{n+2} \in A \,|\, X_1, \ldots, X_n]$$

$\Longrightarrow$ this direct approach has the merit of highlighting that a two–step ahead prediction is essential for getting the whole machinery going, but is of little practical help.

Indirect approach via de Finetti's representation theorem

(1) Define a prior distribution Q on the space $\mathscr{P}$ of probability measures on $\mathbb{X}$ and denote by $\tilde{P}$ the random probability measure whose law is given by Q.

(2) Recall that by de Finetti's representation theorem a sequence of $\mathbb{X}$–valued observations $(X_n)_{n\geq 1}$ is exchangeable if and only if for any $n \geq 1$

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{\mathscr{P}} \prod_{i=1}^{n} P(A_i) Q(\mathrm{d}P)$$

$\Longrightarrow$ Q is the de Finetti measure of an exchangeable sequence $(X_n)_{n\geq 1}$

(3) Derive the distribution of Q conditionally on $X^{(n)} := (X_1, \ldots, X_n)$ aka posterior distribution i.e. $Q(\,\cdot\,|\,X^{(n)}) \Longrightarrow$ in the dominated case the posterior is derived via Bayes theorem, otherwise (e.g. the case of discrete nonparametric priors) it becomes trickier.

(4) Given $Q(\,\cdot\,|\,X^{(n)})$ the predictive distributions are obtained via marginalization. E.g., the one–step predictive distributions are

$$\mathbb{P}[X_{n+1} \in A\,|X^{(n)}] = \int_{\mathscr{P}} P(A)\, Q(\mathrm{d}P|X^{(n)}) = \mathbb{E}[\tilde{P}(A)|X^{(n)}]$$

$\Longrightarrow$ How can one specify a random probability measure $\tilde{P}$ whose distribution Q acts as nonparametric prior?

The DP via finite–dimensional distributions [Ferguson, 1973]

The original definition of the DP was in terms of a consistent family of finite–dimensional Dirichlet distributions.

The Dirichlet distribution, $D_{\boldsymbol{\alpha}}$ with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, is the probability distribution on $\Delta_{k-1} := \{(x_1, \ldots, x_{k-1}) : x_1 \geq 0, \ldots, x_{k-1} \geq 0, \sum_{i=1}^{k-1} x_i \leq 1\}$ given by

$$D_{\boldsymbol{\alpha}}(\mathrm{d}x_1, \ldots, \mathrm{d}x_{k-1}) = \frac{\Gamma(\theta)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \left(\prod_{i=1}^{k-1} x_i^{\alpha_i - 1} \right) \left(1 - \sum_{i=1}^{k-1} x_i \right)^{\alpha_k} \mathrm{d}x_1 \ldots \mathrm{d}x_{k-1}.$$

> *Let α be a finite measure on $\mathbb{X}$. A random probability measure is termed Dirichlet process, $\tilde{\mathscr{D}}_\alpha$ if for any measurable partition $A_1, \ldots, A_k$ of $\mathbb{X}$ and $k \geq 1$ one has*
>
> $$(\tilde{\mathscr{D}}_\alpha(A_1), \ldots, \tilde{\mathscr{D}}_\alpha(A_k)) \sim D_{\boldsymbol{\alpha}} \qquad \text{with} \quad \boldsymbol{\alpha} = (\alpha(A_1), \ldots, \alpha(A_k))$$

$\Longrightarrow$ Decomposition: $\alpha(\,\cdot\,) = \theta\, P^*(\,\cdot\,)$ with $\theta > 0$ and P^* a probability measure.

Generalizable? YES and NO.

$\Longrightarrow$ In order to preserve mathematical tractability one needs distributions on Δ_{k-1} which are "additive" in the shape parameters α_i's. Best example Normalized Inverse Gaussian (N–IG) process [Lijoi, Mena & P, 2005].

The DP via a stick–breaking construction [Sethuraman, 1994]

Stick–breaking provides an alternative DP construction:

Let $(V_i)_{\geq 1}$ be a sequence of i.i.d. random variables, with $V_i \sim B_{1,\theta}$ and define random probability weights $(\tilde{p}_j)_{j\geq 1}$ as

$$\tilde{p}_1 = V_1, \qquad \tilde{p}_j = V_j \prod_{i=1}^{j-1}(1 - V_i) \quad j = 2, 3, \dots . \tag{1}$$

If $(Z_i)_{j\geq 1}$ is a sequence of i.i.d. random variables whose common probability distribution is P^*, then

$$\sum_{j\geq 1} \tilde{p}_j \, \delta_{Z_j}(\,\cdot\,) = \tilde{\mathscr{D}}_{\theta\, P^*(\,\cdot\,)}\,. \tag{2}$$

Generalizable? YES and NO.

- Use independent $V_i \sim B_{1-\sigma,\theta+i\sigma}$ to obtain a PY process with parameters $\sigma \in (0,1)$, $\theta > -\sigma$ and P^* [Perman, Pitman & Yor, 1992].
- The only mathematically tractable $\tilde{P}$ with i.i.d. or independent V_i's are DP and PY processes $\Longrightarrow$ for other "interesting" $\tilde{P}$ the V_i's are dependent and explicit derivation is difficult; a nice example is the N–IG process [Favaro, Lijoi & P, 2012b].

The DP via normalization & neutrality to the right

Yet another two constructions of the Dirichlet process [Ferguson, 1973 & 1974; Doksum, 1974] are based on processes with independent increments:

- Normalization: Let $\{\xi_t : t \geq 0\}$ be a gamma process i.e. a process with independent increments s.t. $\xi_t \sim \text{ga}_{\alpha((0,t]),1}$ with α a finite measure on $\mathbb{R}^+$. Then
$$\tilde{F}(t) = \frac{\xi_t}{\lim_{t\to\infty} \xi_t}$$
is the random cdf associated to a Dirichlet process on $\mathbb{R}^+$.
- Neutrality to the right: Let $\{\zeta_t : t \geq 0\}$ be a "suitable" process with independent increments. Then
$$\tilde{F}(t) = 1 - \mathrm{e}^{-\zeta_t}$$
is the random cdf associated to a Dirichlet process on $\mathbb{R}^+$.

Generalizable? YES:
1. replace the specific independent increments process, whose transformation leads to the DP, with a general independent increments process or better by a completely random measure leading to general classes of prior distributions
2. derive their conditional (posterior) distributions and then obtain the predictive distributions as posterior expected values (which thanks to de Finetti's representation theorem automatically identify an exchangeable

Completely random measures [Kingman, 1967; Kingman, 1993]

Let $\mathbb{M}$ be the space of boundedly finite measures on some Polish space $\mathbb{X}$.

Completely random measure

A random element $\tilde{\mu}$ on $(\Omega, \mathscr{F}, \mathbb{P})$ taking values in $\mathbb{M}$ such that, for any $d \geq 2$ and disjoint sets $A_1, \ldots, A_d$,

$$\tilde{\mu}(A_1), \ldots, \tilde{\mu}(A_d) \qquad \text{are mutually independent}$$

is said to be a completely random measure (CRM) on $\mathbb{X}$.

The realizations of a CRM are a.s. discrete and it can be represented as

$$\tilde{\mu}(\,\cdot\,) = \sum_{i=1}^{\infty} J_i \delta_{Z_i}(\,\cdot\,) + \sum_{j=1}^{M} W_i \delta_{\mathrm{x}_i}(\,\cdot\,)$$

- both the $\mathbb{X}$–valued locations Z_i's and the positive jumps J_i's are random;
- $\mathrm{x}_1, \ldots, \mathrm{x}_M$, with $M \in \{0, 1, \ldots, \infty\}$, are fixed jump points in $\mathbb{X}$ and the (non–negative) random jumps W_j's are mutually independent and independent from $\sum_{i=1}^{\infty} J_i \delta_{Z_i}$.

$\Longrightarrow$ In the following we assume $\tilde{\mu}$ has no fixed point of discontinuity unless otherwise specified.

- A CRM $\tilde{\mu}$ is uniquely characterized by its *Laplace functional*

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}(\mathrm{d}x)}\right] = \mathrm{e}^{-\int_{\mathbb{R}^+\times\mathbb{X}}[1-\mathrm{e}^{-v\,g(x)}]\nu(\mathrm{d}v,\mathrm{d}x)} \qquad (\star)$$

for any $\mathbb{R}^+$–valued $g \in \mathscr{G}_\nu := \{g : \int_{\mathbb{R}^+\times\mathbb{X}}[1-\mathrm{e}^{-v\,g(x)}]\nu(\mathrm{d}v,\mathrm{d}x) < \infty\}$.

In $(\star)$ ν stands for the intensity measure, which characterizes the CRM $\tilde{\mu}$.
In fact, to any measure ν on $\mathbb{R}^+ \times \mathbb{X}$ s.t.

$$\int_B \int_{\mathbb{R}^+} \min\{v,1\}\,\nu(\mathrm{d}v,\mathrm{d}x) < \infty \qquad \text{for any } B \in \mathscr{X}$$

there corresponds a unique CRM $\tilde{\mu}$.

- If $\nu(\mathrm{d}v,\mathrm{d}x) = \rho(\mathrm{d}v)\alpha(\mathrm{d}x)$ the laws of the J_i's and the X_i's are independent and $\tilde{\mu}$ is termed *homogeneous CRM*.
On the other hand, if $\nu(\mathrm{d}v,\mathrm{d}x) = \rho_x(\mathrm{d}v)\alpha(\mathrm{d}x)$, we have a *non–homogeneous CRM*.
$\Longrightarrow$ The terminology "homogeneous" and "non homogeneous" carries over to models derived from CRMs.

Examples of CRMs

Let be α a σ–finite measure, $\sigma \in (0,1)$, $\tau > 0$ and β a strictly positive function.

1. gamma CRM: $\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\mathrm{e}^{-\tau v}}{v}\mathrm{d}v\alpha(\mathrm{d}x)$
2. σ–stable CRM: $\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)\, v^{1+\sigma}}\mathrm{d}v\alpha(\mathrm{d}x)$
3. inv. Gaussian CRM: $\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\mathrm{e}^{-\tau v}}{\sqrt{2\pi}\, v^{3/2}}\mathrm{d}v\alpha(\mathrm{d}x)$
4. gener. gamma CRM: $\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\sigma\, \mathrm{e}^{-\tau v}}{\Gamma(1-\sigma)\, v^{1+\sigma}}\mathrm{d}v\alpha(\mathrm{d}x)$ [Brix, 1999]
5. exten. gamma CRM: $\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\mathrm{e}^{-\beta(x)v}}{v}\mathrm{d}v\alpha(\mathrm{d}x)$ [Dykstra & Laud, 1981]
6. superposed gamma CRM: $\nu(\mathrm{d}v, \mathrm{d}x) = \frac{(1-\mathrm{e}^{-\tau v})}{(1-\mathrm{e}^{-v})}\frac{\mathrm{e}^{-v}}{v}\mathrm{d}v\,\alpha(\mathrm{d}x)$

BNP models based on CRMs

Many BNP models arise as transformations of CRMs:

- Normalized completely random measures [Regazzini, Lijoi and P, 2003]

  $\Longrightarrow$ Special cases: Dirichlet process, N–IG process, Normalized generalized gamma process

- Gibbs–type priors [Gnedin & Pitman, 2006]

  $\Longrightarrow$ Special cases: Dirichlet process, Pitman–Yor process

- Neutral to the right processes [Doksum, 1974]: Let $\tilde{\mu}$ be a CRM on $\mathbb{R}^+$ s.t. $\tilde{\mu}(\mathbb{R}^+) = \infty$. Then

$$\tilde{P}((0,t]) = 1 - \mathrm{e}^{-\tilde{\mu}((0,t])}$$

  is a *neutral to the right process*.

  $\Longrightarrow$ Special case: beta-Stacy process [Walker & Muliere, 1997] if $\tilde{\mu}$ s.t.

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\mathrm{e}^{-v\,c(x)\,P^*((x,\infty))}}{1-\mathrm{e}^{-v}}\, c(x)\,\mathrm{d}v\,P^*(\mathrm{d}x)$$

  with c some strictly positive piecewise continuous function. It reduces to the Dirichlet process if $c(x) \equiv c$.

Instead of specifying $\tilde{P}$, an alternative approach originated in survival analysis consists in focusing on the hazard rate

$$\frac{\mathrm{d}F(x)}{1-F(x-)} = \Pr[\text{death at } (x, x+\mathrm{d}x] \,|\, \text{ survival time } \geq x]$$

or on the cumulative hazard

$$H(x) = \int_0^x \frac{\mathrm{d}F(u)}{1-F(u-)}$$

Note that from the above relation one deduces

$$F(t) = 1 - \mathrm{e}^{-H_c(t)} \prod_{j:t_j \leq t} (1 - H(\{t_j\}) \qquad (\diamond)$$

where H_c is the continuous part of H and the t_j's are the discontinuity points of H. From $(\diamond)$ one immediately sees that the size of the jumps of H have to be less than 1.
If H is absolutely continuous $(\diamond)$ reduces to

$$F(t) = 1 - \mathrm{e}^{-H(t)}.$$

- Random cumulative hazard models [Hjort, 1990]: Let $\tilde{\mu}$ be a CRM on $\mathbb{R}^+$ s.t. $\tilde{\mu}(\mathbb{R}^+) = \infty$ and the jump part of ν is concentrated on $[0, 1]$, *i.e.*

$$\rho_x((1, \infty)) = 0 \qquad \forall x > 0.$$

Then $\tilde{H}(x) := \tilde{\mu}((0, x])$ is a random cumulative hazard CRM model.

Special case: beta process which corresponds to

$$\nu(\mathrm{d}v, \mathrm{d}x) = c(x)\, v^{-1} (1 - v)^{c(x)-1}\, \mathrm{d}v\, \alpha(\mathrm{d}x) \quad \text{for any } 0 < v < 1, x \geq 0$$

with c some strictly positive function.

$\Longrightarrow$ Beta CRM on $\mathbb{X}$ s.t. $\tilde{\mu}(\mathbb{X}) < \infty$ is de Finetti measure of the Indian Buffet process [Griffiths & Ghahramani, 2006; Thibaux & Jordan, 2007].

- Mixture hazard models [James, 2005]: Define a *random mixture hazard* as

$$\tilde{h}(y) = \int_{\mathbb{X}} k(y|x)\, \tilde{\mu}(\mathrm{d}x)$$

where $\tilde{\mu}$ is a CRM and k a kernel s.t. $\tilde{H}(t) = \int_0^t \tilde{h}(y)\mathrm{d}y \to \infty$.

$\Longrightarrow$ Popular kernel choices:

  - increasing hazard rates: $k(y|x) = \mathbb{I}_{(0 \leq x \leq y)}$ [Dykstra & Laud, 1981];
  - decreasing hazard rates: exponential–type kernels;
  - flexible kernels: rectangular kernel $k(y|x) = \mathbb{I}_{(|t-x| \leq \tau)}$ with $\tau > 0$; OU kernel $k(y|x) = \sqrt{2\kappa} \exp\left(-\kappa\,(t - x)\right) \mathbb{I}_{(0 \leq x \leq t)}$ with $\kappa > 0$.

Normalized CRMs [Regazzini, Lijoi & P, 2003]

$\Longrightarrow$ definition of a random probability measure via normalization of a CRM:

(A) $\tilde{\mu}(\mathbb{X})$ almost surely finite $\Longleftrightarrow \int_{\mathbb{R}^+ \times \mathbb{X}} [1 - \mathrm{e}^{-\lambda v}] \, \nu(\mathrm{d}v, \mathrm{d}x) < \infty \quad \forall \lambda > 0$

$[\Longleftrightarrow \alpha(\mathbb{X}) < \infty$ if $\tilde{\mu}$ is a homogeneous CRM$]$

(B) $\tilde{\mu}(\mathbb{X})$ almost surely strictly positive $\Longleftrightarrow \nu(\mathbb{R}^+, B) = \infty$ for some B s.t. $\alpha(B) > 0$ (infinite activity)

$\left[\Longleftrightarrow \rho(\mathbb{R}^+) = \infty \text{ if } \tilde{\mu} \text{ is a homogeneous CRM}\right]$

Definition.
Let $\tilde{\mu}$ be a CRM on $\mathbb{X}$ satisfying (A) and (B). Then the random probability measure on $\mathbb{X}$ given by

$$\tilde{P}(\,\cdot\,) = \frac{\tilde{\mu}(\,\cdot\,)}{\tilde{\mu}(\mathbb{X})}$$

is a normalized completely random measure (NRMI).

$\Longrightarrow$ The acronym stems from its original definition on $\mathbb{R}$ as normalized random measure with independent increments.

A NRMI is uniquely characterized by the intensity ν of the associated CRM $\tilde{\mu}$.

A priori moments: Dirichlet process

The DP is recovered as special case of NRMI by taking $\tilde{\mu}$ as a gamma CRM with α a finite measure on $\mathbb{X} \Longrightarrow \tilde{P} \sim \mathscr{D}_\alpha$ with parameter measure $\alpha = \theta P^*$.

A priori moments

Let $\tilde{P} \sim \mathscr{D}_\alpha$. Then for any $A, B \in \mathscr{X}$

$$\mathbb{E}[\tilde{P}(A)] = \frac{\alpha(A)}{\alpha(\mathbb{X})} = P^*(A)$$

$$\mathrm{Var}(\tilde{P}(A)) = \frac{P^*(A)P^*(A^c)}{\theta + 1}.$$

$$\mathrm{Cov}(\tilde{P}(A), \tilde{P}(B)) = \frac{P^*(A \cap B) - P^*(A)P^*(B)}{\theta + 1}$$

Proofs are straightforward by noting that marginally, for any A, one has $\tilde{P}(A) \sim \mathrm{Be}(\alpha(A), \alpha(A^c))$. E.g. $\mathbb{E}[\tilde{P}(A)] = \frac{\alpha(A)}{\alpha(A)+\alpha(A^c)} = \frac{\alpha(A)}{\alpha(\mathbb{X})} = P^*(A)$.

How do we derive a priori moments for general NRMI since we do not know the marginal of $\tilde{\mu}(A)$ let alone $\tilde{P}(A)$? We know the Laplace transform of $\tilde{\mu}(A)$!

A priori moments: homogeneous NRMI

A priori moments

Let $\tilde{P}$ be a homogeneous NRMI with $\nu(\mathrm{d}v, \mathrm{d}x) = \rho(\mathrm{d}v)\theta P^*(\mathrm{d}x)$. Then for any $A, B \in \mathscr{X}$

$$\mathbb{E}[\tilde{P}(A)] = \frac{\alpha(A)}{\alpha(\mathbb{X})} = P^*(A)$$

$$\mathrm{Var}(\tilde{P}(A)) = P^*(A)P^*(A^c)\ \mathscr{I}_\theta.$$

$$\mathrm{Cov}(\tilde{P}(A), \tilde{P}(B)) = (P^*(A \cap B) - P^*(A)P^*(B))\ \ \mathscr{I}_\theta$$

with $\tau_i(u) = \int_0^\infty v^i \mathrm{e}^{-uv} \rho(\mathrm{d}v)$ for $i = 1, 2 \ldots$ and

$$\mathscr{I}_\theta = \mathbb{P}(X_1 = X_2) = \theta \int_0^\infty u\, \tau_2(u)\, \mathrm{e}^{-\theta \int (1-\mathrm{e}^{-uv})\rho(\mathrm{d}v)} \mathrm{d}u.$$

The crucial thing to note is that from a mathematical point of view, it is best to view normalization as follows:

$$\tilde{P}(A) = \frac{\tilde{\mu}(A)}{\tilde{\mu}(\mathbb{X})} = \tilde{\mu}(A) \int_0^\infty \exp\{-u\, \tilde{\mu}(\mathbb{X})\}\ \mathrm{d}u.$$

Sketch of proof for the first moment:

$$\mathbb{E}\left[\tilde{P}(A)\right] = \mathbb{E}\left[\frac{\tilde{\mu}(A)}{\tilde{\mu}(\mathbb{X})}\right] = \mathbb{E}\left[\int_0^\infty \tilde{\mu}(A)\exp\left\{-u\,\tilde{\mu}(\mathbb{X})\right\}\,\mathrm{d}u\right]$$

$$= \int_0^\infty \mathbb{E}\left[\tilde{\mu}(A)\,\mathrm{e}^{-u\,\tilde{\mu}(A)}\right]\,\mathbb{E}\left[\mathrm{e}^{-u\,\tilde{\mu}(A^c)}\right]\mathrm{d}u \qquad \text{[by Fubini \& independence]}$$

$$= \int_0^\infty \mathbb{E}\left[-\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{e}^{-u\,\tilde{\mu}(A)}\right]\,\mathrm{e}^{-\alpha(A^c)\psi(u)}\mathrm{d}u \qquad \left[\text{with } \psi(u) = \int_0^\infty (1-\mathrm{e}^{-uv})\rho(\mathrm{d}v)\right]$$

$$= \int_0^\infty \left\{-\frac{\mathrm{d}}{\mathrm{d}u}\mathbb{E}\left[\mathrm{e}^{-u\,\tilde{\mu}(A)}\right]\right\}\,\mathrm{e}^{-\alpha(A^c)\psi(u)}\mathrm{d}u$$

$$= \int_0^\infty \left\{-\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{e}^{-\alpha(A)\psi(u)}\right\}\,\mathrm{e}^{-\alpha(A^c)\psi(u)}\mathrm{d}u = \alpha(A)\int_0^\infty \left\{\frac{\mathrm{d}}{\mathrm{d}u}\psi(u)\right\}\mathrm{e}^{-\alpha(\mathbb{X})\psi(u)}\mathrm{d}u$$

$$= \frac{\alpha(A)}{\alpha(\mathbb{X})}\int_0^\infty -\frac{\mathrm{d}}{\mathrm{d}u}\mathrm{e}^{-\alpha(\mathbb{X})\psi(u)}\mathrm{d}u \qquad \text{[having multiplied and divided by } \alpha(\mathbb{X})\text{]}$$

$$= \frac{\alpha(A)}{\alpha(\mathbb{X})}\left(1-\mathrm{e}^{-\alpha(\mathbb{X})\rho(\mathbb{R}^+)}\right) \qquad \left[\text{by construction of NRMI } \rho(\mathbb{R}^+) = \infty\right]$$

$$= \frac{\alpha(A)}{\alpha(\mathbb{X})} = P^*(A)$$

Conjugacy

Dirichlet process:

- Let $\tilde{\mu}$ be a gamma CRM with α a finite measure on $\mathbb{X} \Longrightarrow$ NRMI is a Dirichlet process, $\tilde{\mathscr{D}}_\alpha$ with parameter measure $\alpha = \theta\, P^*$.
- Conjugacy [Ferguson, 1973]: *The posterior distribution of $\tilde{\mathscr{D}}_\alpha$ given a sample $X^{(n)}$ coincides with the distribution of $\tilde{\mathscr{D}}^*_\alpha$ for which*

$$\alpha^*(\,\cdot\,) = \alpha(\,\cdot\,) + \sum_{i=1}^{K_n} N_i \delta_{X_i^*}$$

Other NRMIs:

- *Let $\mathscr{Q}$ be the set of all NRMIs laws and let $\tilde{P} \in \mathscr{Q}$. The posterior distribution of $\tilde{P}$, given a sample $X^{(n)}$, is still in $\mathscr{Q}$ if and only if $\tilde{P}$ is a Dirichlet process.* [James, Lijoi & P, 2006]

  $\Longrightarrow$ Conjugacy is a distinctive feature of DP.
- Nonetheless, conditionally on the data $X^{(n)}$ and a latent variable U_n, the (posterior) distribution of any NRMI $(\tilde{P}|X^{(n)}, U)$ is still in $\mathscr{Q}$.

The latent variable U_n [James, Lijoi & P, 2009]

The latent variable U_n has a precise meaning summarizing the normalization procedure and the distribution of $\tilde{\mu}(\mathbb{X})$.

- Set $\tau_m(u|x) = \int_{\mathbb{R}^+} s^m \, \mathrm{e}^{-us} \, \rho_x(\mathrm{d}s)$ for any $m \geq 1$ and $x \in \mathbb{X}$;
- U_n is a positive r.v. whose density, conditional on $X^{(n)}$, is for any $n \geq 1$

$$f(u|X^{(n)}) \propto u^{n-1} \, \mathrm{e}^{-\psi(u)} \prod_{j=1}^{k} \tau_{n_j}(u|X_j^*)$$

with $\psi(u) := \int_{\mathbb{R}^+ \times \mathbb{X}} [1 - \mathrm{e}^{-uv}] \, \nu(\mathrm{d}v, \mathrm{d}x)$.

Remark: *The distribution of* $(U_n|X^{(n)})$ *is a mixture of gamma distributions*

$$f(u|X^{(n)}) = \int_{(0,+\infty)} \frac{y^n}{\Gamma(n)} \, u^{n-1} \, \mathrm{e}^{-yu} \, \bar{Q}\left(\mathrm{d}y|X^{(n)}\right)$$

where the mixing measure $\bar{Q}(\,\cdot\,|X^{(n)})$ *is the posterior distribution of* $\tilde{\mu}(\mathbb{X})$*, given* $X^{(n)}$.

The posterior characterization [James, Lijoi & P, 2009]

The posterior distribution of $\tilde{\mu}$, given $X^{(n)}$, is a mixture with respect to U_n:

$$\left(\tilde{\mu} \,\middle|\, X^{(n)}, U_n\right) = \tilde{\mu}_u + \sum_{i=1}^{K_n} J_i^{(u)}\, \delta_{X_i^*}$$

where

(i) *$\tilde{\mu}_u$ is a CRM with intensity*

$$\nu^{(u)}(\mathrm{d}x, \mathrm{d}v) = \mathrm{e}^{-uv}\, \rho_x(\mathrm{d}v)\, \alpha(\mathrm{d}x)$$

(ii) *for $i = 1, \ldots, K_n$, the jumps $J_i^{(u)}$ at X_i^* admit density*

$$f_i(s) \propto s^{N_i} \mathrm{e}^{-us} \rho_{X_i^*}(\mathrm{d}s)$$

(iii) *$\tilde{\mu}_u$ and $J_i^{(u)}$ $(i = 1, \ldots, K_n)$ are mutually independent.*

(iv) *Moreover, given $X^{(n)}$ and U_n, $\tilde{P}$ is again a NRMI:*

$$\left(\tilde{P} \,\middle|\, X^{(n)}, U_n\right) = \frac{\tilde{\mu}_u + \sum_{i=1}^{K_n} J_i^{(u)}\, \delta_{X_i^*}}{\tilde{\mu}_u(\mathbb{X}) + \sum_{i=1}^{K_n} J_i^{(u)}}$$

$\Longrightarrow$ The posterior distribution of $\tilde{\mu}$ and of $\tilde{P}$ can be seen as laws, respectively, of an exchangeable random measure and of its normalization.

General proof strategy for posteriors of CRM based models

- In general, CRM based BNP models consist in a suitable transformation of a CRM, say $T(\tilde{\mu})$, e.g. normalization, exponentiation or convolution.
- Instead of tackling directly the derivation of $T(\tilde{\mu})|X^{(n)}$, it is best to obtain first the distribution of $\tilde{\mu}|X^{(n)}$; the distribution of $T(\tilde{\mu})|X^{(n)}$ then typically follows as an immediate consequence.
- The tool to characterize the posterior $\tilde{\mu}|X^{(n)}$ is the Laplace functional i.e. one has to derive

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}(\mathrm{d}x)} \,\middle|\, X^{(n)}\right]$$

- If one is "lucky" (e.g. for NTR models), the posterior turns out to be the sum of two independent components, namely an updated CRM and random jumps at the distinct sample values i.e.

$$\tilde{\mu}^* + \sum_{i=1}^{K_n} J_i \delta_{X_i^*} \qquad \text{with} \quad \tilde{\mu}^* \perp\!\!\!\perp (J_i)_{i\geq 1}$$

which in terms of Laplace functionals corresponds to

$$\mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}(\mathrm{d}x)} \,\middle|\, X^{(n)}\right] = \mathbb{E}\left[\mathrm{e}^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}^*(\mathrm{d}x)}\right] \prod_{i=1}^{K_n} \mathbb{E}\left[\mathrm{e}^{-g(X_i^*)J_i}\right]$$

- In general, "somewhere hidden" inside the posterior distribution $\tilde{\mu}|X^{(n)}$ there is still a discrete random structure of the form

$$\tilde{\mu}^* + \sum_{i=1}^{K_n} J_i \delta_{Z_i^*} \qquad \text{with} \quad \tilde{\mu}^* \perp\!\!\!\perp (J_i)_{i\geq 1} \qquad (*)$$

  but one has to dig it out!

- The strategy consists in identifying a suitable latent structure such that conditionally on it a posterior structure of the form $(*)$ appears.
  For instance, in the NRMI case the identification of the latent variable U_n allows to derive

$$\mathbb{E}\left[e^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}(\mathrm{d}x)} \,\middle|\, X^{(n)}\right] = \int_0^\infty \mathbb{E}\left[e^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}(\mathrm{d}x)} \,\middle|\, X^{(n)}, U_n = u\right] f(u|X^{(n)})\mathrm{d}u$$

$$= \int_0^\infty \mathbb{E}\left[e^{-\int_{\mathbb{X}} g(x)\,\tilde{\mu}_u(\mathrm{d}x)}\right] \prod_{i=1}^{K_n} \mathbb{E}\left[e^{-g(X_i^*)J_i^{(u)}}\right] f(u|X^{(n)})\mathrm{d}u$$

- Importantly, the same strategy applies also beyond the exchangeable framework and the identification of a suitable latent structure allows to uncover posterior CRM structures of the form $(*)$.

The generalized gamma NRMI [Lijoi & P, 2010]

Let $\tilde{\mu}$ be a generalized gamma (GG) CRM with α a finite measure on $\mathbb{X}$ and

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, \mathrm{e}^{-v}\mathrm{d}v\, \alpha(\mathrm{d}x)$$

$\Longrightarrow$ the resulting NRMI is the normalized generalized gamma (NGG) process.

The (posterior) distribution of $\tilde{\mu}$, given $X^{(n)}$ and U_n,

$$\tilde{\mu}_u + \sum_{i=1}^{K_n} J_i^{(u)}\, \delta_{X_i^*}$$

where

(i) $\tilde{\mu}_u$ is a GG–CRM with intensity measure

$$\nu^{(u)}(\mathrm{d}v, \mathrm{d}x) = \tfrac{\sigma}{\Gamma(1-\sigma)}\, v^{-1-\sigma}\, \mathrm{e}^{-(u+1)v}\mathrm{d}v\, \alpha(\mathrm{d}x)$$

(ii) the jump at X_i^* is $J_i \sim \mathrm{Ga}(n_i - \sigma, u+1)$, for $i = 1, \ldots, K_n$;

(iii) $\tilde{\mu}^{(u)}$ and J_i $(i = 1, \ldots, k)$ are mutually independent.

Moreover, the distribution of U_n, conditional on $X^{(n)}$, is

$$f(u|X^{(n)}) \propto \frac{u^{n-1}\, \mathrm{e}^{-\alpha(\mathbb{X})(1+u)^{\sigma}}}{(u+1)^{n-k\sigma}}.$$

NRMI predictive distributions [James, Lijoi & P, 2009]

The predictive distributions are of the form

$$P[X_{n+1} \in \mathrm{d}x_{n+1} | X^{(n)}] = w^{(n)}\, \alpha(\mathrm{d}x_{n+1}) + \frac{1}{n} \sum_{j=1}^{K_n} w_j^{(n)}\, \delta_{X_j^*}(\mathrm{d}x_{n+1})$$

with $\tau_m(u|x) = \int_0^\infty s^m \, \mathrm{e}^{-us}\, \rho_x(\mathrm{d}s)$ for $m = 1, 2, \ldots$ and

$$w^{(n)} = \frac{1}{n} \int_0^{+\infty} u\, \tau_1(u|x_{n+1})\, f(u|X^{(n)})\, \mathrm{d}u \quad w_j^{(n)} = \int_0^{+\infty} u\, \frac{\tau_{N_j+1}(u|X_j^*)}{\tau_{N_j}(u|X_j^*)}\, f(u|X^{(n)})\, \mathrm{d}u.$$

If $\tilde{\mu}$ is a σ-stable CRM, the predictive distributions become

$$\mathbb{P}\left[X_{n+1} \in \cdot \,\middle|\, X^{(n)} \right] = \frac{\sigma K_n}{n}\, P^*(\,\cdot\,) + \frac{n - \sigma K_n}{n} \sum_{i=1}^{K_n} \frac{N_i - \sigma}{n - \sigma K_n} \delta_{X_i^*}(\,\cdot\,).$$

Recap: general class of CRM based random probability measures (whose laws are de Finetti measure of exchangeable sequences) $\Longrightarrow$ derivation of the posterior distribution $\Longrightarrow$ predictive distributions (posterior expected value).

In general: Analogous strategy can be adopted for other models based on CRMs and, importantly, the posteriors enjoy common structural features.

Posterior distribution of NTR processes [Ferguson, 1974]

Recall that a random cdf $\tilde{F}$ is neutral to the right if there exists a CRM $\tilde{\mu}$ such that $\tilde{F}(t) = 1 - \mathrm{e}^{-\tilde{\mu}((0,t])}$.

Posterior representation is again in terms of the CRM:

$$\left(\tilde{\mu}((0,t])|X^{(n)}\right) = \tilde{\mu}^*((0,t]) + \sum_{i=1}^{K_n} J_i \delta_{X_i^*}((0,t])$$

(i) $\tilde{\mu}^*$ is a CRM with intensity

$$\nu^*(\mathrm{d}v) = \mathrm{e}^{-\bar{N}(x)\,v}\,\rho_x(\mathrm{d}v)\alpha(\mathrm{d}x)$$

where $\bar{N}(x) =$ "at risk process" $= \sum_{i=1}^{K_n} N_i \mathbb{I}_{X_i^*}([x,\infty))$

(ii) the distribution of each jump J_i is given in terms of ν and $\bar{N}$

(iii) $\tilde{\mu}^*$ and the random jumps $J_1, \ldots, J_{K_n}$ are mutually independent

$\Longrightarrow$ NTR process priors are conjugate [Doksum, 1974] in the sense that if $\tilde{F}$ is NTR also $\tilde{F}|X^{(n)}$ is NTR (but based on a different CRM)

$\Longrightarrow$ Derivation of predictive distributions quite complicated [James, 2006]

Posterior distribution of mixture hazard rates [James, 2005]

Recall that a mixture hazard model is of the form $\tilde{h}(y) = \int_{\mathbb{X}} k(y|x)\tilde{\mu}(\mathrm{d}x)$ with k a kernel and $\tilde{\mu}$ a CRM. The data $(Y_n)_{n\geq 1}$ are exchangeable with de Finetti measure the law of $\tilde{f}(y) = \tilde{h}(y)\mathrm{e}^{-\tilde{H}(y)}$.

Yet again the posterior representation is expressed in terms of the posterior distribution of the CRM.
Introduce a set of latent variables $X^{(n)}$ and the conditional distribution of the CRM given $Y^{(n)}$, $(\tilde{\mu}\,|\,Y^{(n)})$, is given in terms of $(\tilde{\mu}\,|\,Y^{(n)}, X^{(n)})$ mixed over $(X^{(n)}\,|\,Y^{(n)})$:

$$\left(\tilde{h}(y)|\,Y^{(n)}, X^{(n)}\right) = \int_{\mathbb{X}} k(t|x)\tilde{\mu}^*(\mathrm{d}x) + \sum_{i=1}^{K_n} J_i k(y|X_i^*),$$

where

(i) μ^* is a CRM with intensity

$$\nu^*(\mathrm{d}v, \mathrm{d}x) := \mathrm{e}^{-v\sum_{i=1}^{n}\int_0^{y_i} k(y|x)\mathrm{d}y}\,\rho(\mathrm{d}v|x)\,\alpha(\mathrm{d}x)$$

(ii) the latent variables X_i^*'s are the location of the random jumps J_i's (for $i = 1, \ldots, K_n$), which are, mutually independent and independent of $\tilde{\mu}^*$.

(iii) the distribution of $X^{(n)}|\,Y^{(n)}$ is also available.

Pitman–Yor (PY) process

Take a generalized gamma (GG) CRM with finite base measure $\alpha := \theta\, P^*$ s.t.

$$\nu(\mathrm{d}v, \mathrm{d}x) = \frac{\sigma\, \mathrm{e}^{-v}}{\Gamma(1-\sigma)\, v^{1+\sigma}} \theta P^*(\mathrm{d}x) \qquad \sigma \in (0,1)$$

Assign a $\mathrm{ga}(\theta/\sigma, 1)$ prior to θ and denote the mixed GG CRM by $\tilde{\varphi}$.
$\Longrightarrow$ $\tilde{P}(\,\cdot\,) = \frac{\tilde{\varphi}(\,\cdot\,)}{\tilde{\varphi}(\mathbb{X})}$ is a PY process. [Pitman & Yor, 1997]

Posterior characterization of the PY process [Lijoi & P, 2010]

Given $X^{(n)}$ and a latent U_n, the (posterior) distribution of $\tilde{\varphi}$ coincides with

$$\tilde{\mu}_u + \sum_{i=1}^{K_n} J_i^{(u)}\, \delta_{X_i^*}$$

where

(i) $\tilde{\mu}_u$ is a GG–CRM with intensity $\nu^{(u)}(\mathrm{d}v) = \frac{\sigma}{\Gamma(1-\sigma)}\, s^{-1-\sigma}\, \mathrm{e}^{-u\,v} \mathrm{d}v$

(ii) the jumps $J_i^{(u)} \sim \mathrm{ga}(N_i - \sigma, u)$

(iii) $\tilde{\mu}_u$ and $J_i^{(u)}$ $(i = 1, \ldots, K_n)$ are mutually independent.

The density of $U_n | X^{(n)}$ is $f(u|X^{(n)}) \propto u^{\theta + K_n\sigma - 1}\, \mathrm{e}^{-u^\sigma}$

$\Longrightarrow$ The PY and NGG processes have the same posterior structure except for a different latent variable U_n.

Posterior characterization of the PY process [Pitman, 1996]

From the previous posterior representation, one obtains also the simpler one:

$$\tilde{P} \,|\, X^{(n)} = (1 - \sum_{i=1}^{K_n} w_i)\, \tilde{P}^{(K_n)} + \sum_{j=1}^{K_n} w_j\, \delta_{X_j^*}$$

where $\tilde{P}^{(K_n)}$ *is a PY process with parameters* σ *and* $\theta + \sigma K_n$ *and*

$$(w_1, \ldots, w_{K_n}) \sim Dir(N_1 - \sigma, \ldots, N_{K_n} - \sigma, \theta + \sigma K_n).$$

From the posterior representations one then easily obtains the simple predictive distributions associated to the PY process:

$$\mathbb{P}\left[X_{n+1} \in \cdot \,\middle|\, X^{(n)}\right] = \frac{\theta + \sigma K_n}{\theta + n}\, P^*(\,\cdot\,) + \frac{n - \sigma K_n}{\theta + n}\, \frac{\sum_{i=1}^{K_n} (N_i - \sigma)\delta_{X_i^*}(\,\cdot\,)}{n - \sigma K_n}.$$

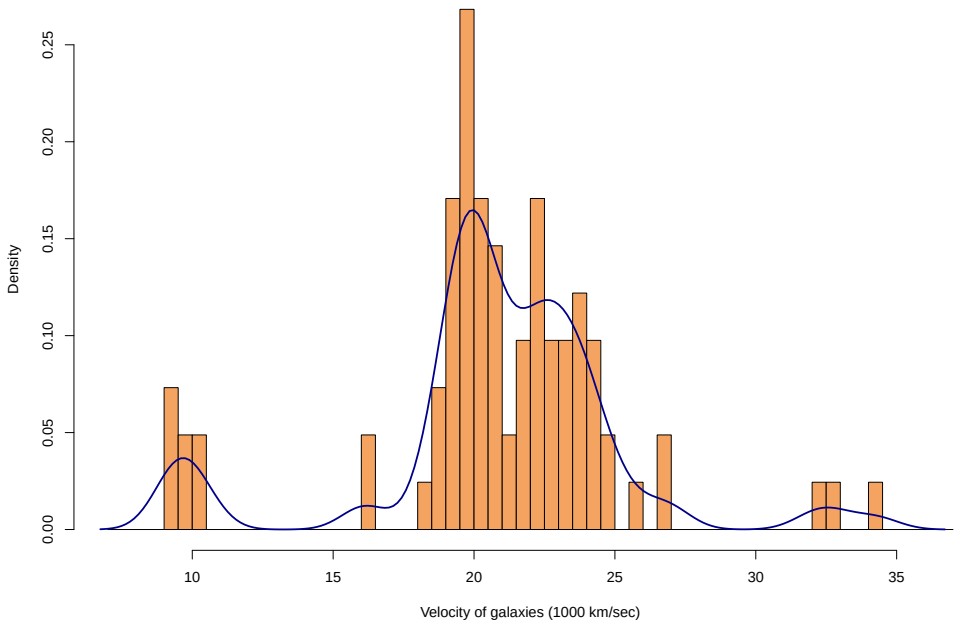
$\Longrightarrow$ if $\sigma = 0$, the PY reduces to the Dirichlet process and $\frac{\theta + K_n \sigma}{\theta + n}$ to $\frac{\theta}{\theta + n}$.

Mixture models for density estimation and clustering

Popular use of discrete $\tilde{P} = \sum_{j \geq 1} \tilde{p}_j \, \delta_{Z_j}$ on $\mathbb{X}$ is at the latent level of a mixture

$$\text{Random density} = \tilde{f}(y) = \int_{\mathbb{X}} f(y \mid x) \, \tilde{P}(\mathrm{d}x) = \sum_{j \geq 1} \tilde{p}_j f(y \mid Z_j)$$

with f a kernel (e.g. Gaussian) such that $\int_{\mathbb{X}} f(y \mid x) \, \mathrm{d}y = 1$ for any $x \in \mathbb{X}$.



$\Longrightarrow$ Special cases: Mixture of Dirichlet process [Lo, 1984], mixture of PY process [Ishwaran & James, 2001], mixture of N–IG and normalized generalized gamma process [Lijoi, Mena & P, 2005, 2007]

Truncated NRMIs [Barrios et al. (2013)]

- **Marginal algorithms:** integrate out $\tilde{P}$ and evaluate $\mathbb{E}[\tilde{f} \mid \text{data}]$ through a Gibbs sampler
- **Conditional algorithms:**
  - The posterior distribution of $\tilde{\mu}$, given $X^{(n)}$ and U_n is a CRM:

$$\left(\tilde{\mu} \mid X^{(n)}, U_n\right) = \sum_{i=1}^{\infty} J_i \delta_{Z_i} + \sum_{i=1}^{K_n} J_i^{(u)} \delta_{X_i^*}$$

  - representation due to Ferguson and Klass (1972)

$$J_1 > J_2 > \cdots$$

  - simulate the truncated NRMI and re-normalize the weights

$$\tilde{P}_{H,\text{tr}} = \sum_{i=1}^{H} \tilde{p}_i^* \, \delta_{Z_i} + \sum_{i=1}^{K_n} \tilde{\omega}_i^* \, \delta_{X_i^*}$$

Extensions and refinements can be found, e.g., in Arbel & P. (2017), Griffin (2016) and Argiento et al. (2016).

Pros & cons

Pros

- Ordering of the J_i's ensures that the truncation procedure retains most relevant random probability masses
- Free from limitations of Pólya urn–type marginal algorithms

Cons

- Simulation of the ordered weights J_i in the truncated representation may lead to numerical issues

As with stick–breaking priors:

- Are there efficient alternatives, that work also for non–exchangeable data?
- For NRMIs the literature on priors on the finite-dimensional simplex is limited to the Dirichlet multinomial process

The Dirichlet multinomial prior

- A classical finite-dimensional prior is the **Dirichlet multinomial** process

$$\tilde{P}_H = \sum_{h=1}^{H} \tilde{p}_h \delta_{Z_h}, \quad (\tilde{p}_1, \ldots, \tilde{p}_{H-1}) \sim D_{\boldsymbol{\alpha}}, \quad \boldsymbol{\alpha} = (\theta/H, \ldots, \theta/H), \quad Z_h \overset{\text{iid}}{\sim} P^*$$

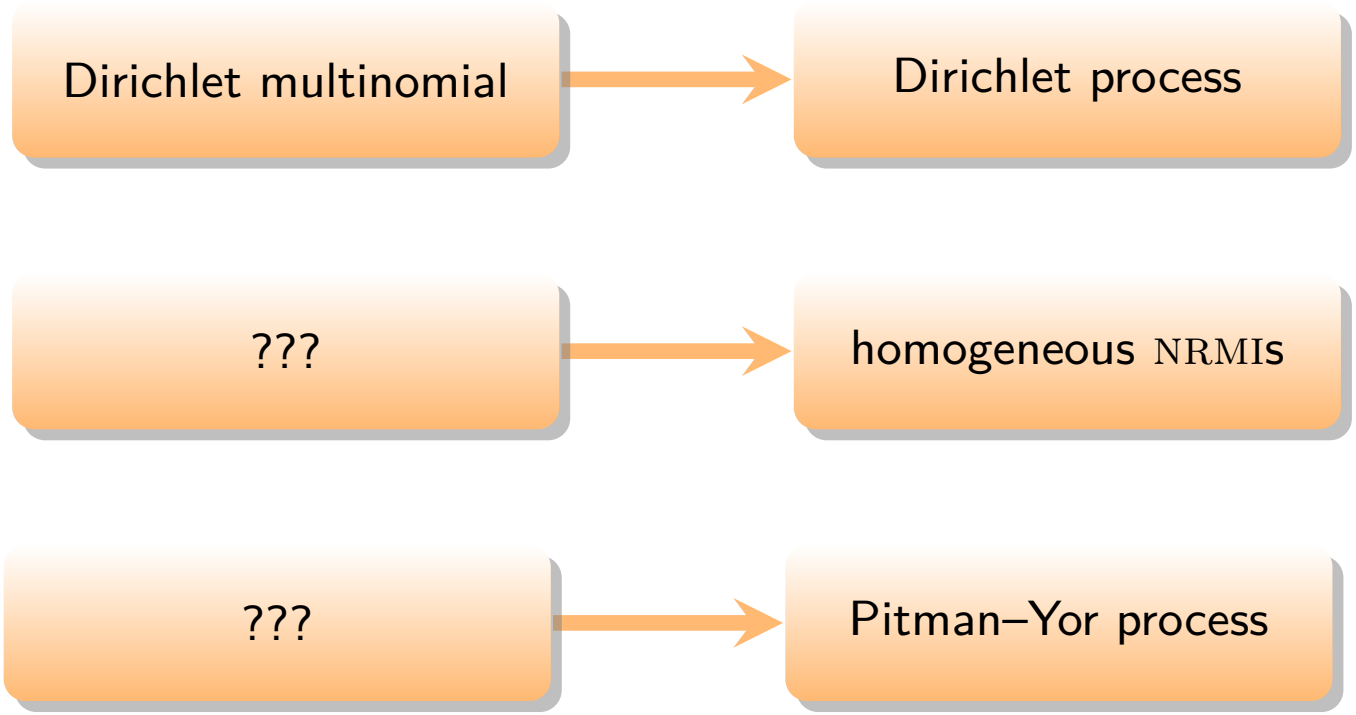
- The Dirichlet multinomial $\tilde{P}_H$ and the Dirichlet process $\tilde{P} \sim \mathscr{D}_{\theta P^*}$ are related through

$$\mathcal{L}_{\tilde{P}_H} \Longrightarrow \mathcal{L}_{\tilde{P}}, \qquad H \to \infty.$$

- This motivates the use of $\tilde{P}_H$ also as an approximation of $\tilde{P}$.

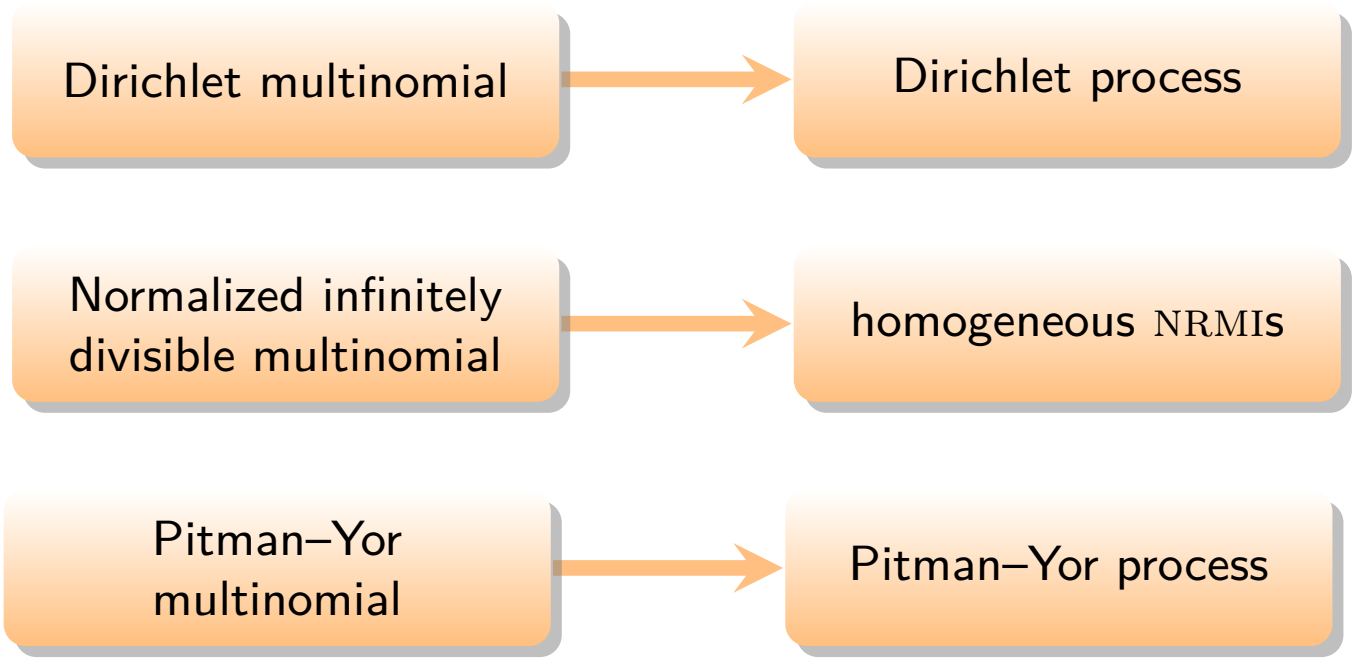
- However, the connection between $\tilde{P}_H$ and $\tilde{P}$ is far deeper, in fact holding for any finite value H.

Goal



- Natural extension of the Dirichlet multinomial?
- Would they preserve a good degree of analytical and computational tractability?

Goal



- Natural extension of the Dirichlet multinomial?
- Would they preserve a good degree of analytical and computational tractability? Yes!

NID multinomial processes

Normalized infinitely divisible multinomial processes

Let $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ be such that $\int_0^\infty \rho(s)\,\mathrm{d}s = \infty$. The random probability measure $\tilde{P}_H$ is a normalized infinitely divisible multinomial process if

$$(\tilde{P}_H \mid \tilde{P}_H^*) \sim \mathrm{NRMI}(\theta, \rho; \tilde{P}_H^*), \qquad \tilde{P}_H^* = \frac{1}{H}\sum_{h=1}^{H} \delta_{Z_h}, \qquad Z_h \overset{\text{iid}}{\sim} P^*$$

In symbols $\tilde{P}_H \sim \mathrm{NIDM}_H(\theta, \rho; P^*)$

- It can be described in terms of infinitely divisible random variables

$$\mathbb{E}\,\mathrm{e}^{-\lambda J} = \exp\left\{ -\theta \int_0^\infty \left(1 - \mathrm{e}^{-\lambda s}\right) \rho(s)\,\mathrm{d}s \right\}$$

$J \sim \mathrm{ID}(\theta, \rho)$

NID multinomial processes: a characterization

Characterization

Let $\rho : \mathbb{R}^+ \to \mathbb{R}^+$ be such that $\int_0^\infty \rho(s)\,\mathrm{d}s = \infty$ and

$$J_h \overset{\text{iid}}{\sim} \text{ID}\left(\frac{\theta}{H}, \rho\right).$$

If $\tilde{P}_H \sim \text{NIDM}_H(\theta, \rho; P^*)$ then

$$\tilde{P}_H = \frac{\tilde{\mu}_H}{\tilde{\mu}_H(\mathbb{X})} = \sum_{h=1}^{H} \frac{J_h}{J^*}\,\delta_{Z_h} \qquad J^* = \sum_{h=1}^{H} J_h = \tilde{\mu}_H(\mathbb{X})$$

Moreover, the vector of probability weights is referred to as *normalized infinitely divisible*

$$(\tilde{p}_1, \ldots, \tilde{p}_{H-1}) = \left(\frac{J_1}{J^*}, \ldots, \frac{J_{H-1}}{J^*}\right) \sim \text{NID}\left(\frac{\theta}{H}, \ldots, \frac{\theta}{H}; \rho\right)$$

Multinomial processes

- The baseline $\mathbb{E}[\tilde{P}_H \mid \tilde{P}_H^*] = \tilde{P}_H^*$ is atomic
- This causes serious analytical difficulties, compared to the case where $\tilde{P}_H^*$ is non–atomic.

Additional example: the PY *multinomial process*

Let $\sigma \in [0, 1)$ and $\theta > -\sigma$. The random probability measure $\tilde{P}_H$ is a Pitman–Yor multinomial process if

$$(\tilde{P}_H \mid \tilde{P}_H^*) \sim \mathrm{PY}(\sigma, \theta; \tilde{P}_H^*), \qquad \tilde{P}_H^* = \frac{1}{H} \sum_{h=1}^{H} \delta_{Z_h}, \qquad Z_h \overset{\text{iid}}{\sim} P^*$$

In symbols $\tilde{P}_H \sim \mathrm{PYM}_H(\sigma, \alpha; P^*)$

Weak limits, as $H \nearrow \infty$

Weak limit for NIDM processes [Lijoi, P. & Rigon, 2019]

Let $\tilde{P}_H \sim \text{NIDM}_H(\theta, \rho; P^*)$ and $\tilde{P} \sim \text{NRMI}(\theta, \rho; P^*)$. Then

$$\mathcal{L}_{\tilde{P}_H} \implies \mathcal{L}_{\tilde{P}} \qquad H \nearrow \infty$$

Weak limit for PYM processes [Lijoi, P. & Rigon, 2020]

Let $\tilde{P}_H \sim \text{PYM}_H(\sigma, \theta; P^*)$ and $\tilde{P} \sim \text{PY}(\sigma, \theta; P^*)$. Then

$$\mathcal{L}_{\tilde{P}_H} \implies \mathcal{L}_{\tilde{P}} \qquad H \nearrow \infty$$

- Distributional properties hold for any choice of H, not just at the limit.

Summary of the results

EPPF of multinomial processes

EPPF of a NIDM process [Lijoi, P. & Rigon, 2019]

The EPPF associated to a $\tilde{P}_H \sim \text{NIDM}_H(\theta, \rho; P^*)$, with P^* diffuse, is

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} \frac{1}{\Gamma(n)} \int_{\mathbb{R}^+} u^{n-1} \mathrm{e}^{-\theta\psi(u)} \left\{ \prod_{j=1}^{k} \mathcal{V}_{n_j}(u) \right\} \mathrm{d}u,$$

with $\psi(u) = \int_{\mathbb{R}^+} (1 - \mathrm{e}^{-us}) \rho(s)\, \mathrm{d}s$ and $\mathcal{V}_m(u) = \left\{ (-1)^m \frac{\partial^m}{\partial u^m} \mathrm{e}^{-\frac{\theta}{H}\psi(u)} \right\} \mathrm{e}^{\frac{\theta}{H}\psi(u)}$

EPPF of a PYM process [Lijoi, P. & Rigon, 2020]

The EPPF associated to a $\tilde{P}_H \sim \text{PYM}_H(\sigma, \theta; P^*)$, with P^* diffuse, is

$$\Pi_H(n_1, \ldots, n_k) = \frac{H!}{(H-k)!} \frac{1}{(\theta+1)_{n-1}} \sum_{(\ell_1, \ldots, \ell_k)} \frac{\Gamma(\theta/\sigma + |\boldsymbol{\ell}^{(k)}|)}{\sigma \Gamma(\theta/\sigma + 1)} \prod_{j=1}^{k} \frac{\mathscr{C}(n_j, \ell_j; \sigma)}{H^{\ell_j}},$$

with $\boldsymbol{\ell}^{(k)} = (\ell_1, \ldots, \ell_k) \in \times_{j=1}^{k} \{1, \ldots, n_j\}$, and $\mathscr{C}(n, k; \sigma)$ the generalized factorial coefficient.

Distribution of $K_{n,H}$

Once the EPPF $\Pi(n_1, \ldots, n_k)$ is available, one can determine

- The distribution of the number of clusters, or partitions sets, $K_{n,H}$
- The predictive distribution of X_{n+1}, conditional on $X^{(n)} = (X_1, \ldots, X_n)$
- The posterior distribution of $\tilde{P}$, given $X^{(n)}$

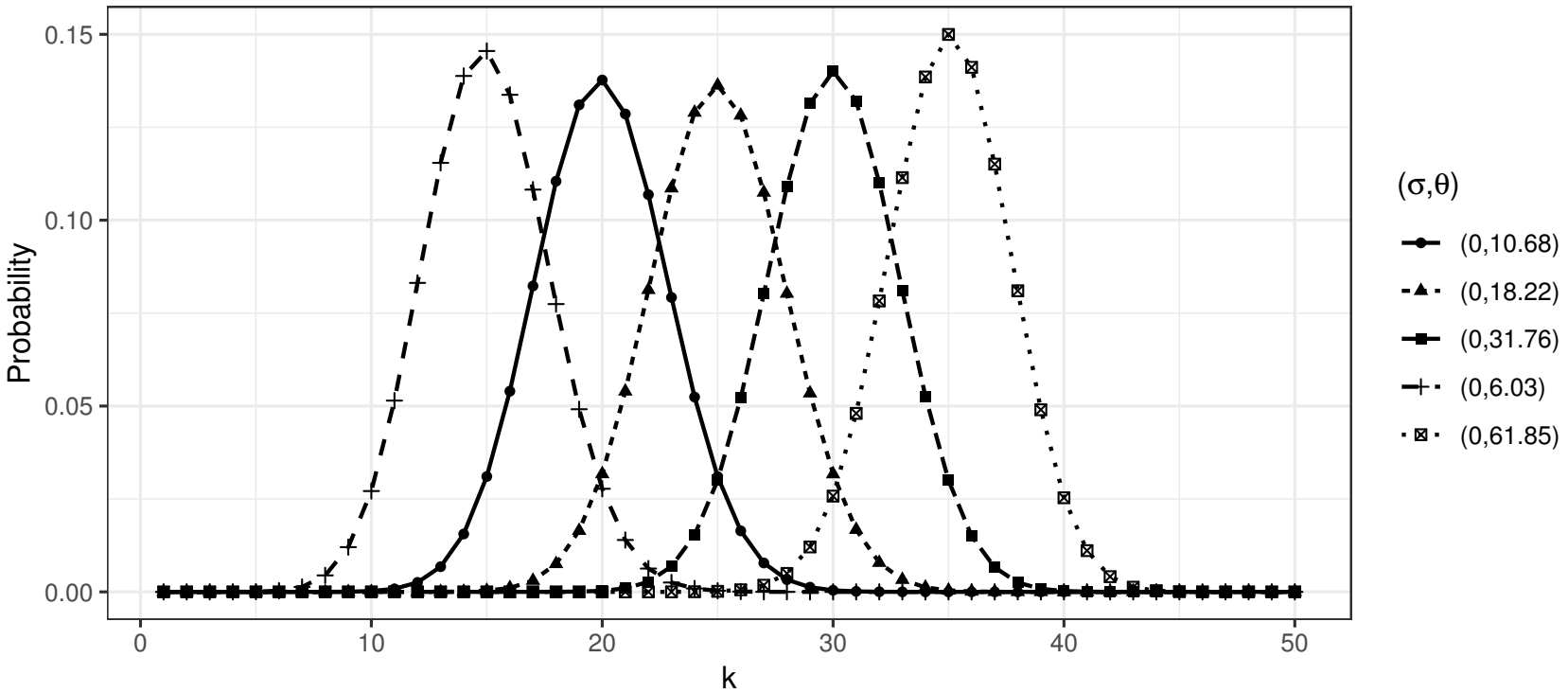
Distribution of $K_{n,H}$ [Lijoi, P. & Rigon, 2019]

If $X_1, \ldots, X_n \,|\, \tilde{P}_H \overset{\text{iid}}{\sim} \tilde{P}_H$ and $\tilde{P}_H \sim \text{NIDM}_H(\theta, \rho; P^*)$, the number $K_{n,H}$ of distinct values in $X^{(n)}$ has probability distribution

$$\mathbb{P}(K_{n,H} = k) = \frac{H!}{H^k (H-k)!} \sum_{\ell=0}^{n-k} \frac{1}{H^\ell} \mathscr{S}(\ell + k, k) \mathbb{P}(K_{n,\infty} = \ell + k),$$

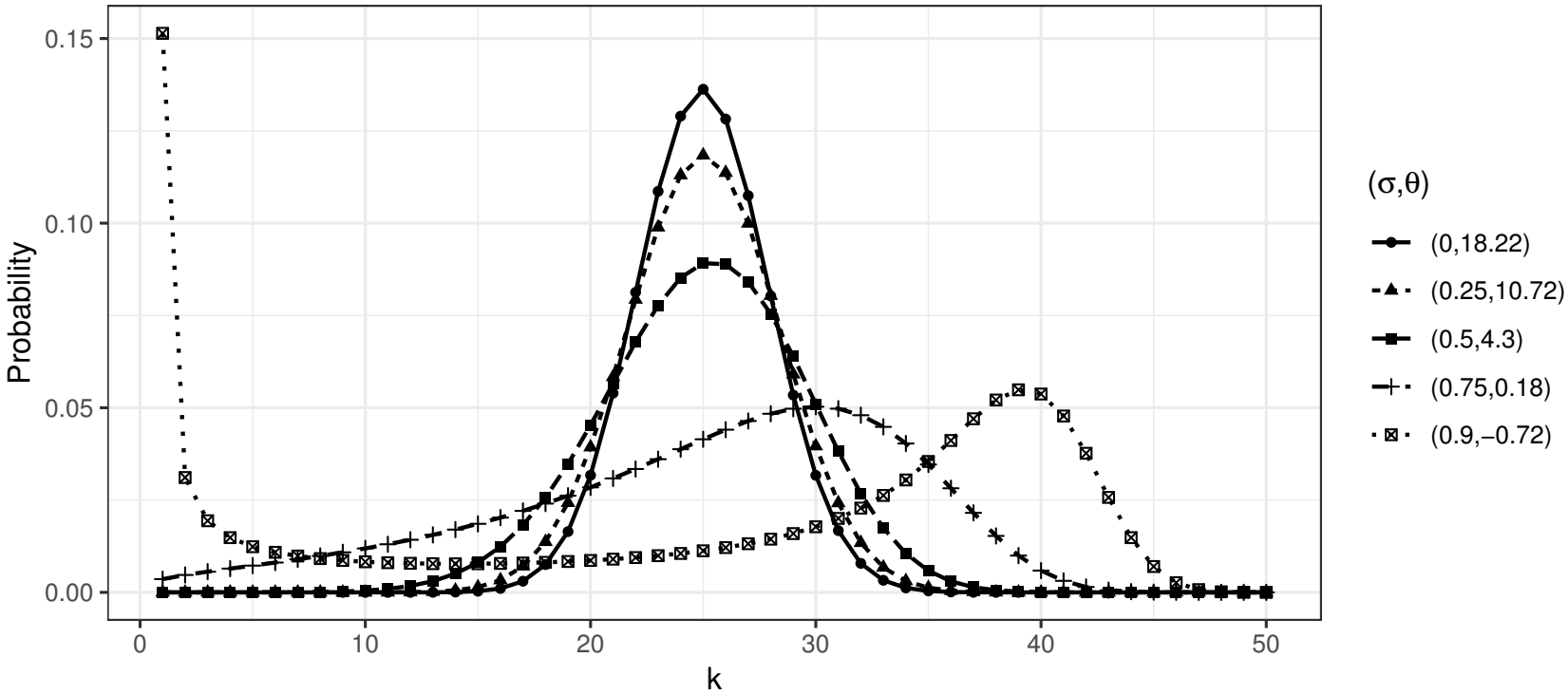
$k \leq \min\{H, n\}$ and $\mathscr{S}(\ell, k)$ is the Stirling number of the second kind.

An example with the Dirichlet multinomial



- Distribution of the number of cluster when $n = 100$, $H = 50$ in the Dirichlet multinomial case ($\sigma = 0$).
- The Dirichlet multinomial is highly informative.

An example with the PYM, with $\sigma > 0$



- Distribution of the number of cluster when $n = 100$, $H = 50$, and for various choices of (σ, θ) so that $\mathbb{E}(K_{n,H}) = 25$.

- More robust specification $\Longrightarrow$ De Blasi et al. (2015); Canale and P. (2017).

Urn schemes for NIDM processes

- Define

$$\Delta_{m,H}(u) = \sum_{\ell=1}^{m} \left(\frac{\theta}{H}\right)^{\ell-1} \frac{1}{\ell!} \sum_{\boldsymbol{q}} \binom{m}{q_1 \ \cdots \ q_\ell} \prod_{r=1}^{\ell} \int_0^\infty s^{q_r} \, \mathrm{e}^{-us} \rho(s) \, \mathrm{d}s,$$

  over all vectors $\boldsymbol{q} = (q_1, \ldots, q_\ell)$ of positive integers s.t. $\sum_{r=1}^{\ell} q_r = m$.

- The density of the latent random variable $(U_{n,H} \mid X^{(n)})$ is

$$f_H(u \mid X^{(n)}) \propto u^{n-1} \mathrm{e}^{-c\psi(u)} \prod_{j=1}^{k} \Delta_{n_j,H}(u).$$

Predictive distribution [Lijoi, P. & Rigon, 2019]

$$\mathbb{P}(X_{n+1} \in A \mid X^{(n)}) = \left(1 - \frac{k}{H}\right) \frac{\theta}{n} \int_{\mathbb{R}^+} u \Delta_{1,H}(u) f_H(u \mid X^{(n)}) \mathrm{d}u \, P^*(A) +$$
$$+ \sum_{j=1}^{k} \frac{1}{n} \int_{\mathbb{R}^+} u \frac{\Delta_{n_j+1,H}(u)}{\Delta_{n_j,H}(u)} f_H(u \mid X^{(n)}) \mathrm{d}u \delta_{X_j^*}(A).$$

Posterior characterization of a NIDM

The posterior distribution of $\tilde{P}_H = \tilde{\mu}_H/\tilde{\mu}(\mathbb{X}) \sim \text{NIDM}_H(\theta, \rho; P^*)$, conditional on $X^{(n)}$, is a mixture with respect to $U_{n,H}$, with

$$(\tilde{\mu}_H \mid X^{(n)}, U_{n,H}) \stackrel{\text{d}}{=} \sum_{j=k+1}^{H} J_j \delta_{Z_j} + \sum_{j=1}^{k} (J_j + J_j^{(u)}) \delta_{X_j^*},$$

where $Z_{k+1}, \ldots, Z_H$ are iid draws from P^* and

(i) *the jumps* $(J_h \mid X^{(n)}, U_{n,H})$ *for* $h = 1, \ldots, H$ *are iid* $\text{ID}\left(\theta/H, \rho^{(u)}\right)$ *r.v. with*

$$\rho^{(u)}(s) = \text{e}^{-Us}\rho(s)$$

(ii) *the jumps* $(J_j^{(u)} \mid X^{(n)}, U_{n,H})$ *for* $j = 1, \ldots, K_n$ *are independent and nonnegative r.v. such that*

$$\mathbb{E}\left(\text{e}^{-\lambda J_j^{(u)}} \,\middle|\, \boldsymbol{\theta}^{(n)}, U\right) = \Delta_{n_j,H}(\lambda + U)/\Delta_{n_j,H}(U)$$

(iii) $(J_h \mid X^{(n)}, U_{n,H})$ *and* $(J_j^{(u)} \mid X^{(n)}, U_{n,H})$ *are mutually independent.*

What's left to do in the exchangeable case?

What's next? Beyond exchangeability!

From de Finetti (1938):

> *But the case of exchangeability can only be considered as a limiting case: the case in which this "analogy" is, in a certain sense, absolute for all events under consideration. [..] To get from the case of exchangeability to other cases which are more general but still tractable, we must take up the case where we still encounter "analogies" among the events under consideration, but without attaining the limiting case of exchangeability.*

In applications dependence structures more general than exchangeability are required. We focus on data collected under different experimental conditions s.t.

- Homogeneity within each experimental condition
- Heterogeneity across different experimental conditions

Examples: Topic modeling, Meta-Analysis, two-sample problems, nonparametric regression (covariate–indexed data), time dependent data, change-point problems ...

Some References

• Arbel & Prünster, I. (2017). A moment-matching Ferguson & Klass algortihm. *Stat. Comput.* **27**, 3-17.

• Argiento, Bianchini & Guglielmi (2016). A blocked Gibbs sampler for NGG-mixture models via a priori truncation. *Stat. Comput.* **26**, 641–661.

• Brix (1999). Generalized gamma measures and shot-noise Cox processes. *Adv. Appl. Probab.* **31**, 929–953.

• Canale & Prünster (2017). Robustifying Bayesian nonparametric mixtures for count data. *Biometrics* **73**, 174-184.

• De Blasi, Favaro, Lijoi, Mena, Prünster & Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE TPAMI* **37** 212-229.

• Doksum (1974). Tailfree and neutral random probabilities and their posterior distributions. *Ann. Probab.* **2**, 183–201.

• Dykstra & Laud (1981). A Bayesian nonparametric approach to reliability. *Ann. Statist.* **9**, 356-367.

• Ferguson (1973). A Bayesian analysis of some nonp. probl. *Ann. Stat.* **1**, 209-30.

• Ferguson (1974). Prior distributions on spaces of prob. meas. *Ann. Stat.* **2**, 615-29.

• de Finetti (1938). Sur la condition d'équivalence partielle. *Act.sci. industr.* **739**, 5-18

• Fortini, Ladelli, Regazzini (2000). Exchangeability, predictive distributions and parametric models. Sankhya A **62**, 86-109.

• Gnedin & Pitman (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci. (N.Y.)* **138**, 5674-85.

• Griffin (2016). An adaptive truncation method for inference in Bayesian nonparametric models. *Stat. Comput.* **26**, 423–441.
• Griffiths & Ghahramani (2006). Infinite Latent Feature Models and the Indian Buffet Process. *NIPS Proceedings* **18**, 475-482.
• Hjort (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18**, 1259–1294.
• Ishwaran & James (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Stat. Assoc.* **96**, 161-173.
• James (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *Ann. Statist.* **33**, 1771-1799.
• James (2006). Poisson calculus for spatial NTR processes. *Ann. Stat.* **34**, 416-440.
• James, Lijoi & Prünster (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scand. J. Statist.* **33**, 105-120.
• James, Lijoi & Prünster (2009). Posterior analysis for normalized random measures with independent increments. *Scand. J. Statist.* **36**, 76-97.
• Kingman (1967). Completely random measures. *Pacific J. Math.* **21**, 59–78.
• Kingman (1993). *Poisson processes.* Oxford University Press, Oxford.
• Lijoi, Mena & Prünster (2005). Hierarchical mixture modelling with normalized inverse Gaussian priors. *J. Amer. Statist. Assoc.* **100**, 1278-1291.
• Lijoi, Mena & Prünster (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *J. Roy. Statist. Soc. Ser. B* **69**, 715-740.
• Lijoi, Prünster & Rigon (2019). Finite-dimensional Discrete Random Structures and Bayesian Clustering. Submitted.

• Lijoi, Prünster & Rigon (2020). The Pitman-Yor multinomial model for mixture modelling. *Biometrika*, forthcoming.
• Lo (1984). On a class of Bayesian nonparametric estimates. *Ann. Stat.* **12**, 351-57.
• Perman, Pitman & Yor (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92**, 21-39.
• Pitman (1996). Some developments of the Blackwell-MacQueen urn scheme. *IMS Lecture Notes Monogr.* **30**, 245-267.
• Pitman & Yor (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *Ann. Probab.* **25**, 855–900.
• Regazzini, Lijoi & Prünster (2003). Distributional results for means of random measures with independent increments. *Ann. Statist.* **31**, 560-585.
• Walker & Muliere (1997). Beta-Stacy processes and a generalization of the Pólya-urn scheme. *Ann. Statist.* **25**, 1762–1780.